

NOTES 10: SAMPLING DISTRIBUTIONS

Stat 120 | Fall 2025

Prof Amanda Luby

“Big picture” picture:

Quantity	Statistic	Parameter
Mean		
Proportion		
Standard Deviation		
Correlation		
Regression Coefficient		

Carleton publishes an “at a glance” page with some facts and figures about the student body:
<https://www.carleton.edu/about/carleton-at-a-glance/>

Some highlights:

- Geographic distribution:
 - Midwest 35%
 - West 24%
 - South 12%
 - Middle States 11%
 - International/Other 11%
 - New England 7%
- 9% report two or more races
- 16% are among the first generation in their families to attend college
- 65% graduated in the top 10% of their high school class
- 75% are involved in community service

In a moment, we’re going to do a poll to find one of these quantities for our class. Before we do, what is your best guess for each of these quantities?

Example: In this set-up, what is the:

- Population
- Sample

- Parameter
- Statistic

We know that our class will likely not have exactly 35% from the Midwest, but we probably wouldn't expect it to be 0% or 90%.

Sampling variability

We might start to ask ourselves, what if a different set of 32 students enrolled in this course?

First, we create a population.

```
# A tibble: 2,007 x 2
  student_id midwest
    <int> <chr>
1      261 Yes
2      650 Yes
3     1070 No
4      568 Yes
5      108 Yes
6      457 Yes
7      911 No
8     1387 No
9     1471 No
10     1318 No
# i 1,997 more rows
```

Then, we take a random sample:

```
set.seed(100424)
sample1 = carls %>%
  sample_n(32)
sample1
```

```
# A tibble: 32 x 2
  student_id midwest
    <int> <chr>
1     1618 No
2      672 Yes
3      348 Yes
4       55 Yes
5     1822 No
6      363 Yes
7     1649 No
```

```

8      1071 No
9      543 Yes
10     956 No
# i 22 more rows

```

and calculate the proportion of “yes” responses:

```

sample1 %>%
  group_by(midwest) %>%
  summarize(
    n = n()
  ) %>%
  mutate(p_hat = n/sum(n))

```

```

# A tibble: 2 x 3
  midwest      n p_hat
  <chr>    <int> <dbl>
1 No         17 0.531
2 Yes        15 0.469

```

This isn’t super useful, but if we do it a bunch of times, we can start to see what a range of possible samples could look like. (Note: this code requires the `infer` package)

```

many_samples = carls %>%
  rep_sample_n(35, reps = 1000, replace = TRUE) %>%
  group_by(replicate, midwest) %>%
  summarize(
    n = n()
  ) %>%
  mutate(p_hat = n/sum(n)) %>%
  filter(midwest == "Yes")

many_samples

```

```

# A tibble: 1,000 x 4
# Groups:   replicate [1,000]
  replicate midwest      n p_hat
    <int> <chr>    <int> <dbl>
1         1 Yes        10 0.286
2         2 Yes        15 0.429
3         3 Yes         8 0.229
4         4 Yes        12 0.343
5         5 Yes        14 0.4
6         6 Yes        12 0.343
7         7 Yes        11 0.314
8         8 Yes        12 0.343

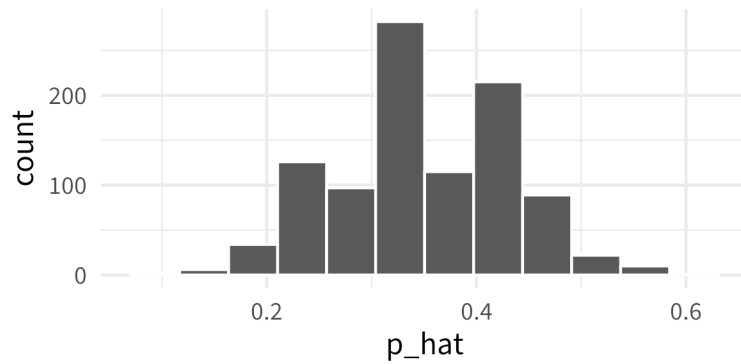
```

```

9          9 Yes      12 0.343
10         10 Yes     14 0.4
# i 990 more rows

```

Looking at this first few rows, we can start to get a sense of the range of possible sample proportions, but there are 990 rows that we can't see. Let's make a graph!



Example: Carleton Mission Statement

In your own words: provide explanations for:

Population distribution

Sample distribution

Sampling distribution